CCLM: Class-Conditional Label Noise Modelling

Albert Tatjer[1], Bhalaji Nagarajan[1(✉)], Ricardo Marques[1], and Petia Radeva[1,2]

[1] Dept. de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain
{acatalta11@alumnes.ub.edu, bhalaji.nagarajan,ricardo.marques,petia.ivanova}@ub.edu

[2] Computer Vision Center, Cerdanyola, Barcelona, Spain

Abstract. The performance of deep neural networks highly depends on the quality and volume of the training data. However, cost-effective labelling processes such as crowdsourcing and web crawling often lead to data with noisy (i.e., wrong) labels. Making models robust to this label noise is thus of prime importance. A common approach is using loss distributions to model the label noise. However, the robustness of these methods highly depends on the accuracy of the division of training set into clean and noisy samples. In this work, we dive in this research direction highlighting the existing problem of treating this distribution globally and propose a class-conditional approach to split the clean and noisy samples. We apply our approach to the popular DivideMix algorithm and show how the local treatment fares better with respect to the global treatment of loss distribution. We validate our hypothesis on two popular benchmark datasets and show substantial improvements over the baseline experiments. We further analyze the effectiveness of the proposal using two different metrics - Noise Division Accuracy and Classiness.

Keywords: Learning with Noisy Labels · Label Noise Modelling · Class-conditional data splitting

1 Introduction

Deep Neural Networks (DNNs) have gained immense attention in the research community due to their success in various challenging domains. High-performing models in all verticals require good quality data of huge volume. However, it is difficult to create such large datasets with high precision as it is labour-intensive, both in collecting and labelling the samples [16]. Crowdsourcing and using semi-automatic labelling pipelines on web-scrapped data reduce the cost of creating datasets. However, they lead to the introduction of noise in the assigned labels [23,24,30]. As a result, real-world datasets tend to have significant label noise,

A. Tatjer and B. Nagarajan—Joint First Authors.
P. Radeva—IAPR Fellow.
R. Marques—Serra Húnter Fellow.

A. Pertusa et al. (Eds.): IbPRIA 2023, LNCS 14062, pp. 3–14, 2023.
https://doi.org/10.1007/978-3-031-36616-1_1

estimated to be in the range of 8.0% to 38.5% [26]. Modern DNNs have a high number of learnable parameters compared to the size of the dataset and often result in overfitting to the label noise [32].

Learning with Noisy Labels (LNL) was introduced in the late 1980s s [1] and has been a long-studied problem with a focus on making the models robust to the label noise. Loss modification-based methods use noise distribution to create more robust loss functions [19], replacing the cross entropy loss or using a noise transition matrix [25]. Correctly estimating the noise and creating methods that are robust to high noise levels are very challenging tasks. Other methods that use sample selection [18] or reweighting [33] use *ad-hoc* criteria to select noisy samples and reduce the impact on the learning process. The challenge here is the selection of a good criterion to split the clean and noise samples.

Semi-supervised learning (SSL) is a helpful strategy for reducing the cost of annotating datasets. SSL-based frameworks use a limited subset of the original dataset being carefully labelled and learn the unlabelled data, which is multifold larger. Recent LNL methods aimed at combining SSL methods and robust loss methods. DivideMix [15] is a popular benchmark that uses two networks to learn the noisy data. At each iteration, one of the networks divides the data into a clean and a noisy datasets, which are then used to train the other network. DivideMix formed the basis for a new family of algorithms [13,27,34]. Several methods based on DivideMix have improved various stages of the pipeline. One of the important questions in most of these methods is *How to effectively demarcate the clean and the noisy samples?* As can be seen with all the works, the selection of clean samples plays an important role in the subsequent sub-tasks. The usage of loss distribution to split the samples has been a well-documented technique. However, not all samples in the training set are learned in the same rate, which hinders noise detection using this technique [21]. Here, we explore this global loss conundrum. We study the division in detail and propose a per-class label noise modelling. We adapt the noise model for each class and create a local threshold used to separate the clean and noisy samples. In order to further study this effect, we propose new metrics to identify the class behaviour of the LNL algorithm. The key contributions of this paper are as follows:

- First, we highlight and study the noise modelling obstacle present in the existing LNL algorithms.
- Second, we propose a class-conditional label noise modelling approach. By replacing the existing global-noise modelling in the popular DivideMix algorithm, we show how the class-conditional approach benefits the algorithm.
- Finally, we introduce new metrics to understand the importance of learning with class importance and show the improvement using two public datasets.

2 Related Works

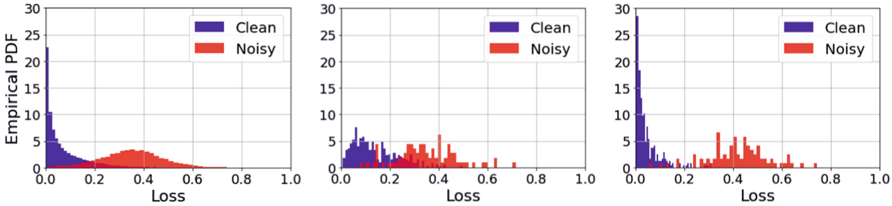
Typical approaches in LNL algorithms include the use of advanced techniques such as sample selection, loss correction, label correction, sample reweighting or using robust loss methods [8]. Sample selection methods deal with finding clean samples using different strategies such as small loss [9] and topological information [31]. Loss correction is achieved by reweighting the loss to avoid overfitting on noisy samples, which helps mitigate the errors introduced by wrongly labelled samples. Similarly, in terms of sample reweighting, the samples that are noisiest are identified and weighted less compared to the clean samples [18]. Another family of LNL algorithms attempt to correct the noise (i.e., errors) in the labels, typically resorting to the notion of *transition matrix*. [7,11,25]. Most of these approaches are based on the assumption that there is a single transition probability between the noisy label and the ground-truth label [10]. Several noise tolerance loss functions have also been proposed to replace the conventional cross-entropy loss [19,20,35]. Recent studies on LNL have shown that hybrid approaches can further boost the algorithms and make them more robust towards label noise [6]. A comprehensive survey on the various LNL algorithms is presented in [26].

Co-teaching [9] of two networks by iteratively selecting clean samples (training samples having small loss) from the other network proved to be very successful. DivideMix [15] used two independent networks for sample selection and then adopted MixMatch [3] to further boost the label correction. DivideMix is a popular benchmarking algorithm in LNL problems. Augmentation plays an important role in LNL problems and was studied over DivideMix [22]. C2D further explored the warmup stage of DivideMix [34]. Probabilistic Noise Prediction [27] used two networks, one to predict the category label and the other to predict the noise type. ProMix [29] used an iterative selection process to select the clean samples using a high-confidence selection technique. SplitNet [13] used an additional learnable module to assist in splitting the clean and noisy samples.

Importance-weighted risk minimization [4] has been studied and employed in different DNNs over the years. A common scheme of importance weighting is weighting the loss terms [12] between different learning tasks. Class-conditional approaches were adapted to LNL algorithms. Generalized Data Weighting via Class-level Gradient Manipulation [5] converted the loss gradient into class-level gradients by applying the chain rule and reweighting the gradients separately. Noise transition matrix [33] was used to capture the class conditional label corruption and used total variation regularization to create more distinguishable predictions. Class-dependent softmax classifier [17] was used to entangle the classification of multi-class features. Our proposed approach is motivated by the LNL methods that deal with the separation of clean and noisy data using different means. As we observe during our study, using a global approach to all the samples is suboptimal and using a class-conditional split at the early stages of the algorithm makes the models more robust towards label noise.

3 The Label Noise Modelling Obstacle

Algorithms which allow efficient learning in the presence of noisy labels often require the division of the training dataset between clean and noisy samples [15]. The samples are then treated differently by the learning algorithm, to extract



(a) All classes. (b) Labelled *otter*. (c) Labelled *road*.

Fig. 1. Empirical per-sample loss distribution between the clean samples and the noisy samples after the warm-up phase (CIFAR-100 20% sym. noise).

the maximum amount of information possible from each of the partitions of the data via semi-supervised learning [3]. The correct division between clean and noisy data is thus crucial for effective learning in the presence of noisy labels. Indeed, if too many noisy labels are classified as clean, then the algorithm risks trusting wrongly labelled data, eventually introducing important biases in the training process and leading to poor results. Conversely, if too many clean samples are deemed noisy, the algorithm would risk disposing of important information, hence potentially impairing the learning outcome. A common approach to splitting the dataset into clean and noisy samples is to create a noise model of the *whole* training data, allowing to characterize the probability of a given sample being clean (i.e., correctly labelled). Given the obvious limitations, it is common to rely on some proxy measure such as loss, which can be used to estimate to which extent a sample is likely to be clean or noisy [2]. We provide a more detailed analysis of this approach and show how the global noise model is suboptimal for some particular classes.

Global Label Noise Modelling. The typical approach to the label noise detection problem relies on the samples' loss to determine whether a given sample is deemed noisy or clean. The sample loss is obtained using a NN model, and the underlying principle is that the NN will tend to yield larger loss values to noisy samples because wrongly labelled data are a more complex pattern and therefore are more difficult to learn [28]. Given a sample x in dataset D, its corresponding one-hot encoded label y and a model p_θ parameterized with the set of parameters θ, the individual loss l of sample x is defined using the cross-entropy loss $l\left(p_\theta(x), y\right)$. The distribution of the cross-entropy loss is shown in Fig. 1a. The distribution appears to follow two modes, with the clean samples (in blue) exhibiting a characteristic loss generally lower than the noisy samples (in red). To turn the cross-entropy loss into a probability of a given sample being correctly (or wrongly) labelled, Li et al. [15] fit a two-component 1-D Gaussian Mixture model (GMM) to the loss distribution l. They select the GMM component g with lower mean and use it to model the probability ω_i of a given sample x_i with loss l_i to be clean. Finally, they use a threshold τ on ω_i to partition the dataset into a clean set $\mathcal{C}_D$ and a noisy set $\mathcal{N}_D$, given by:

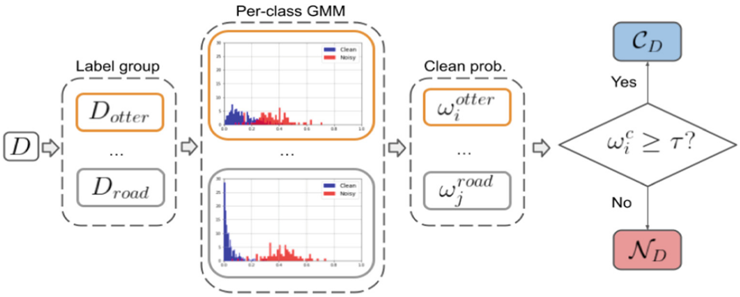


Fig. 2. CCLM splits the dataset according to labels, fitting a GMM to each independent set and yielding the Clean-Noise partition from a decision boundary on each label group's probability distribution.

$$\begin{aligned} \mathcal{C}_D &= \{(x_i, y_i) |\, \omega_i \geq \tau\}_{(x_i,y_i)\in D} \\ \mathcal{N}_D &= \{(x_i, y_i)\,|\, \omega_i < \tau\}_{(x_i,y_i)\in D}\,, \end{aligned} \tag{1}$$

where $\omega_i = p(g|l_i)$ and $l_i = l(p_\theta(x_i), y_i)$.

Limitations of Global Label Noise Modelling: The global noise modelling approach described above assumes that all samples are i.i.d and, therefore, that their per-sample loss tends to behave consistently. However, with the presence of noisy labels, it is obvious that some classes might be harder to learn than others, leading to potentially different characteristic losses across different classes. This situation can be appreciated in Fig. 1, where the per-sample loss distribution for the 'otter' class (Fig. 1b) differs from that of the 'road' class (Fig. 1c) and, more importantly, from the global per-sample loss distribution of the whole dataset (Fig. 1a). We identify this difference between the loss distribution of a particular class and the global loss distribution as an obstacle towards the correct partition of the data into a clean set, $\mathcal{C}_D$ and a noisy set, $\mathcal{N}_D$ which, in its turn, might hinder the training of NN models in the presence of noisy labels. Moreover, we conjecture that taking a class-conditional approach to the problem of label noise detection can lead to better dataset partitions, and thus to superior model accuracy after training in the presence of noisy labels. In the following, we propose a class-conditional approach to the problem of noise label detection, aimed at overcoming the aforementioned limitations of global label noise modelling.

4 Proposed Methodology

Typical approaches in LNL assume all samples to be i.i.d. Consequentially, label noise modelling is performed leveraging solely per-sample loss information, which can impair the accuracy of the trained model. We propose an alternative label-noise model we call Class-Conditional Local Noise Model (CCLM) that allows

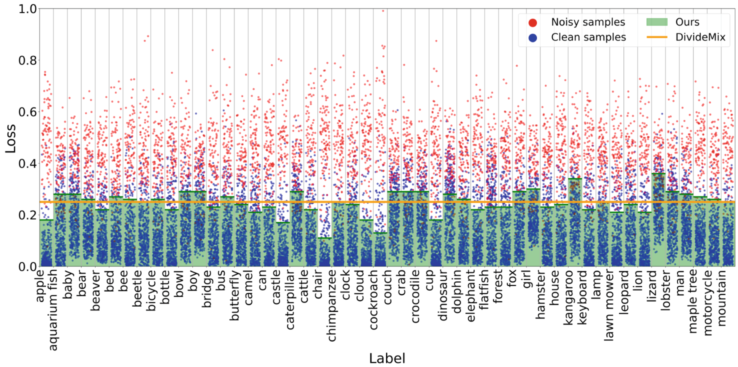


Fig. 3. Per-sample loss distribution of 50 of the CIFAR-100 label groups (20% sym.). Label groups are formed according to the potential noisy labels.

to locally adapt the noise model to the specificity of each considered class, hence improving the overall accuracy of the noisy label detection task. To achieve this, we explicitly include the label class information in the noise modelling process. Our proposed approach is depicted in Fig. 2. We take as input the training set D composed of pairs $(x_i, y_i)_{i=1}^{N}$, N being the total number of samples in D. Then, the dataset is split by sample label y, yielding C disjoint subsets of D such that $D = \cup_{c=1}^{C} D_c$, with C being the total number of possible classes. Then, for each subset D_c, we fit a two-component GMM to the per-sample loss and use the GMM component with a lower mean to model the probability of the samples being clean. We use the threshold τ over the probability of a sample being clean to split the data into noisy and clean within each class. In other words, the probability of a sample being clean is conditional on its loss and its label, $\omega_i^{y_i} = p(g|l_i, y_i)$. Finally, the Clean-Noisy partition is given by:

$$\begin{aligned} \mathcal{C}_D &= \cup_{c=1}^{C} \mathcal{C}_{D_c} = \cup_{c=1}^{C} \{(x_i, y_i) \mid \omega_i^c \geq \tau\}_{(x_i, y_i) \in D_c} \\ \mathcal{N}_D &= \cup_{c=1}^{C} \mathcal{N}_{D_c} = \cup_{c=1}^{C} \{(x_i, y_i) \mid \omega_i^c < \tau\}_{(x_i, y_i) \in D_c}, \end{aligned} \tag{2}$$

Figure 3 illustrates the result of our noise detection model on 20% noise of the CIFAR-100 dataset. The y-axis represents the per-sample loss associated with each sample, with noisy samples being represented in red, and clean samples in blue. The orange horizontal line depicts the split between noisy and clean labels using the typical global noise modelling approach. The green area depicts the per-class split between noisy and clean labels using our CCLM. We clearly see that the shape of the per-sample loss distribution varies across classes (e.g., 'apple' vs. 'aquarium fish'). Moreover, it can be seen that the global noise approach does not distinguish between classes, yielding a single decision boundary for all classes independently of their particular loss features. In contrast, our CCLM provides a locally adapted division between clean and noisy data which allows improving both the split accuracy and the NN accuracy when trained in the presence of noisy labels.

Table 1. Best and Last accuracy on CIFAR-10 and CIFAR-100.

Method		CIFAR-10					CIFAR-100				
		20%	50%	80%	90%	Asym. 40%	20%	50%	80%	90%	Asym. 40%
DivideMix [15]	Best	96.1	94.6	93.2	76	**93.4**	77.3	74.6	60.2	31.5	72.2**
	Last	95.7	94.4	92.9	75.4	92.1	76.9	74.2	59.6	31.0	72.4**
CCLM (ours)	Best	**96.5**	**95.6**	**93.7**	**83.6**	92.6	**79.5**	**76.4**	**61.1**	**33.5**	**75.4**
	Last	96.3	95.3	93.6	82.4	91.5	79.1	75.9	60.9	33.0	75.1

** Results reported from Contrast2Divide [34]. DivideMix does not report this setting.

5 Experiments and Results

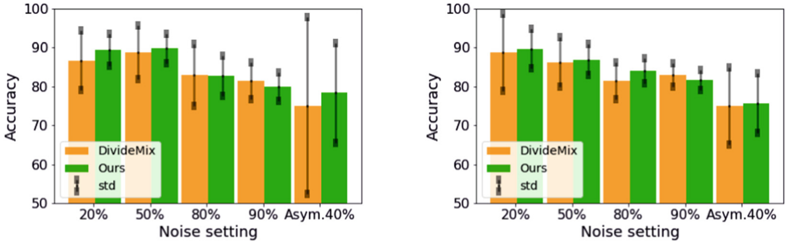
In this section, we present the datasets and the implementation details for evaluating our proposal. We show the performance with two types of label noise - symmetric and asymmetric and compare them against the baseline DivideMix.

5.1 Datasets and Implementation Details

We conduct our evaluations using two benchmark datasets - CIFAR-10 and CIFAR-100 [14]. Both CIFAR datasets contain images of 32×32 RGB pixels, with 50k training samples and 10k test samples. To maintain consistency across the benchmark methods, we follow the same train/test split [15,34]. We conduct our experiments on two different noise types. *Symmetric* (sym.) noise is generated by replacing the labels for a percentage of the training data with a uniform distribution over all the possible labels. *Asymmetric* (asym.) noise is injected by replacing labels with similar classes (e.g. deer $\rightarrow$ horse, dog $\leftrightarrow$ cat). For both datasets, we use 18-layer PreAct Resnet following the settings of DivideMix [15]. We keep all the hyperparameters the same as in DivideMix. We use a batch size of 128 and train the models for 300 epochs using SGD optimizer with a momentum of 0.9 and weight decay of 0.0005. We set an initial learning rate of 0.02 and reduce it by a factor of 10 after 150 epochs. For CIFAR-10, we use a warmup of 10 epochs and for CIFAR-100, we use a warmup of 30 epochs. We perform all experiments using PyTorch on NVIDIA RTX2080Ti GPU. Following previous works [15,34], we report the best test accuracy across all epochs and the average test accuracy over the last 10 epochs (identified as Last).

5.2 Results

First, we show the overall performance of the CCLM against DivideMix, which follows a global noise model approach. We compare it against DivideMix as it is considered a benchmark in LNL. The improvements brought by our proposed method in terms of label noise detection are likely to have a positive impact on algorithms which rely on the partition of the data between clean and noisy.

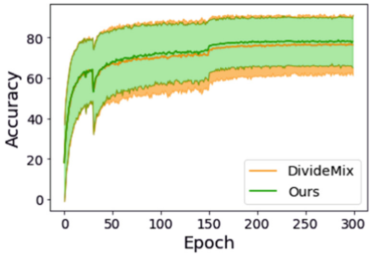


(a) CIFAR-10. (b) CIFAR-100.

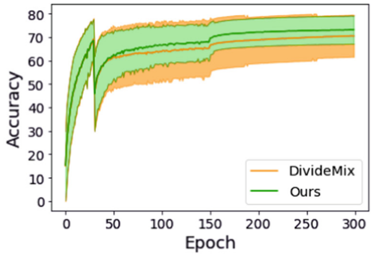
Fig. 4. Clean-Noise split accuracy and std. dev. at the first epoch after warm-up.

Performance. Table 1 shows the test accuracy on CIFAR-10 and CIFAR-100 with different levels of label noise. Our CCLM outperforms the global noise label modelling approach of DivideMix in all the noise settings on CIFAR-100. On CIFAR-10, our approach fares better in all the noise settings except for the asym. noise setting. This superior performance is explained by the local adjustment (i.e., per class) of the label noise model, yielding a more accurate split between clean and noisy samples, hence improving the final model accuracy. Regarding CIFAR-10 asym. noise, where label groups are characterized by a limited number of classes, adopting a class conditional strategy may lead to overly confident discrimination between clean and noisy samples. This phenomenon could be attributed to a lack of related classes, resulting in insignificant "help" on the label group. The results in Table 1 show that our CCLM consistently outperforms the original global noise modelling used in DivideMix, being also more stable at low-noise settings. For both approaches, the accuracy of the models decreases considerably as the noise ratio increases. One of the reasons for it is the over-fitting of the models in large noise settings which might make loss-based noise detection models less efficient. Hence, we further analyse the behaviour of CCLM using two different metrics - Noise Division Accuracy and Classiness.

Noise Division Accuracy. We evaluate the noise detection accuracy at the first epoch after the warmup. To this end, we first collect the per-sample loss provided by the pre-trained DL model. Then, we feed the collected loss values to both the global (DivideMix, using the threshold τ proposed for each noise level) and our CCLM noise detection approaches, yielding two distinct clean-noisy splits. Finally, using the reference solution, we evaluate the accuracy of the splits. The result is shown in Fig. 4, where the average split accuracy and the corresponding standard deviation for CIFAR-10 (Fig. 4a) and CIFAR-100 (Fig. 4b) are depicted, averaged over two runs. Our proposed method achieves a better average accuracy in all noise levels except for 90% sym. noise. It should be noted that a more accurate split after the warm-up does not guarantee better overall model accuracy. In the case of 90% noise, predicting all the samples as noisy would be very bad. The results also show that our method consistently yields a smaller standard deviation than the original one, which indicates that



(a) Test set.



(b) Train set.

Fig. 5. Model accuracy and *Classiness* on CIFAR-100 with 20% sym. noise. The accuracy drop is a result of the ending of the warm-up phase (at epoch 30).

Table 2. Training and Testing Classiness of CIFAR-100.

		Train				Test			
Noise Rate		20%	50%	80%	90%	20%	50%	80%	90%
DivideMix [15]	Average	11.0	14.3	18.2	18.1	16.0	16.7	20.0	19.8
	Best	9	12.7	17.6	18.8	14	15	19	**21**
Ours	Average	7.7	10.9	17.5	17.1	13.6	14.9	19.4	20.2
	Best	**6**	**8.8**	**17.1**	**17.1**	**12**	**13**	**19**	22

our proposal yields a clean-noisy split that performs more evenly over the different classes. This results in a reduced bias towards the easier classes.

Classiness. In the classification task, it is clear that some classes can be more robust to label noise than others. A desired property of any classification model is to achieve similar accuracy in all the classes. In order to capture this accuracy distribution over classes, we introduce the notion of *classiness*, measured as the standard deviation of the model accuracy over the different latent classes. A lower classiness value thus represents a more evenly accurate network. We study classiness during training and testing. Training classiness allows us to inspect whether the training process is biased towards some of the classes. With class-agnostic label-noise modelling, *difficult* classes would be treated as mostly noise. Testing classiness is the desired model property informing about the achieved evenness. We show the training and testing model accuracy and classiness evolution during the training process using 20% sym. noise of CIFAR-100 in Fig. 5, where the first 30 epochs correspond to the warm-up phase. The training classiness (Fig. 5b (painted area)) is smaller on average for our proposed CCLM. This indicates that the model is less biased during training, allowing for better learning in the posterior epochs. The average accuracy of the baseline model and our proposal are similar, but the smaller classiness tends towards better learning seen in the later epochs. Table 2 shows the average training classiness

Table 3. Impact of the clean noise split threshold on the final model accuracy.

τ	0.5	0.6	0.7	0.8	0.9
DivideMix [15]	30.3	**31.5**	31.6	27.6	27.6
Ours	**30.6**	31.1	**32.6**	**33.5**	**30.8**

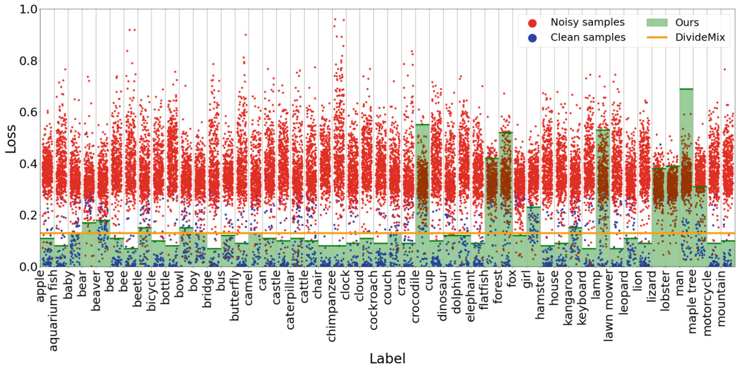


Fig. 6. Per-sample loss distribution of 50 classes of CIFAR-100 (90% sym.).

on CIFAR-100 with all different noise levels and the test classiness of the best epoch. The best epoch is chosen as the one with the better test accuracy. We see that our method has reasonably better test and train classiness on low noise settings and comparable test and train classiness on higher noise settings.

Effect of Clean-Noise Splitting Threshold. In current LNL practice, the label noise modelling step is most commonly used to perform the Clean-Noise split. However, the resulting split also depends on the threshold used as a decision boundary. A critical aspect of the analysis entails examining the impact of the selected threshold on the split outcome. To this end, we present Fig. 6 and Fig. 3, which illustrate the feature space we aim to split. The label-noise modelling procedure yields the probability of a sample belonging to the clean set, $\omega_i = p(g|l_i)$. Following Eq. (1) and Eq. (2), the Clean-Noisy partition is obtained in a sample level by $\omega_i > \tau$. In Table 3, we compare the model accuracy between the global LNL model, and our CCLM, across different split thresholds. Upon inspecting Fig. 6, we observe that a class-agnostic label noise modelling procedure, with a high threshold, disregards harder classes entirely. In contrast, our CCLM is not susceptible to this particular pitfall and is therefore capable of managing higher thresholds.

6 Conclusions

Learning with Noisy Labels is a very important data-centric Machine learning problem. Modern algorithms rely on the division of clean and noisy samples to make further learning decisions. In this paper, we study this division of training samples and show how the global division is sub-optimal. On this front, we propose a class-conditional label noise model, which uses a local division of clean and noisy samples. We validate our approach comparing it with the baseline on two benchmarking datasets. We further introduce two metrics - noise division accuracy and classiness to show the improvements. Our future work focuses on studying the impact of global vs. local threshold in other parts of the algorithm, using other metrics to understand this division and validating the method on other large LNL datasets.

Acknowledgements. This work was partially funded by the Horizon EU project MUSAE (No. 01070421), 2021-SGR-01094 (AGAUR), Icrea Academia'2022 (Generalitat de Catalunya), Robo STEAM (2022-1-BG01-KA220-VET-000089434, Erasmus+ EU), DeepSense (ACE053/22/000029, ACCIÓ), DeepFoodVol (AEI-MICINN, PDC-2022-133642-I00) and CERCA Programme/Generalitat de Catalunya. B. Nagarajan acknowledges the support of FPI Becas, MICINN, Spain. We acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPUs. As Serra Húnter Fellow, Ricardo Marques acknowledges the support of the Serra Húnter Programme.

References

1. Angluin, D., Laird, P.: Learning from noisy examples. Mach. Learn. **2**, 343–370 (1988)
2. Arazo, E., Ortego, D., Albert, P., O'Connor, N., McGuinness, K.: Unsupervised label noise modeling and loss correction. In: International Conference on Machine Learning, pp. 312–321. PMLR (2019)
3. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: a holistic approach to semi-supervised learning. In: NIPS, vol. 32 (2019)
4. Byrd, J., Lipton, Z.: What is the effect of importance weighting in deep learning? In: International Conference on Machine Learning, pp. 872–881. PMLR (2019)
5. Chen, C., et al.: Generalized data weighting via class-level gradient manipulation. In: NIPS, vol. 34, pp. 14097–14109 (2021)
6. Chen, Z., Song, A., Wang, Y., Huang, X., Kong, Y.: A noise rate estimation method for image classification with label noise. In: Journal of Physics: Conference Series, vol. 2433, p. 012039. IOP Publishing (2023)
7. Cheng, D., et al.: Instance-dependent label-noise learning with manifold-regularized transition matrix estimation. In: CVPR, pp. 16630–16639 (2022)
8. Ding, K., Shu, J., Meng, D., Xu, Z.: Improve noise tolerance of robust loss via noise-awareness. arXiv preprint arXiv:2301.07306 (2023)
9. Han, B., et al.: Co-teaching: robust training of deep neural networks with extremely noisy labels. In: NIPS, vol. 31 (2018)
10. Han, J., Luo, P., Wang, X.: Deep self-learning from noisy labels. In: ICCV, pp. 5138–5147 (2019)
11. Hendrycks, D., Mazeika, M., Wilson, D., Gimpel, K.: Using trusted data to train deep networks on labels corrupted by severe noise. In: NIPS, vol. 31 (2018)
12. Khetan, A., Lipton, Z.C., Anandkumar, A.: Learning from noisy singly-labeled data. arXiv preprint arXiv:1712.04577 (2017)
13. Kim, D., Ryoo, K., Cho, H., Kim, S.: SplitNet: learnable clean-noisy label splitting for learning with noisy labels. arXiv preprint arXiv:2211.11753 (2022)

14. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
15. Li, J., Socher, R., Hoi, S.C.: DivideMix: learning with noisy labels as semi-supervised learning. arXiv preprint arXiv:2002.07394 (2020)
16. Liao, Y.H., Kar, A., Fidler, S.: Towards good practices for efficiently annotating large-scale image classification datasets. In: CVPR, pp. 4350–4359 (2021)
17. Liu, S., Zhu, Z., Qu, Q., You, C.: Robust training under label noise by over-parameterization. In: ICML, pp. 14153–14172. PMLR (2022)
18. Liu, X., Luo, S., Pan, L.: Robust boosting via self-sampling. Knowl.-Based Syst. **193**, 105424 (2020)
19. Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S., Bailey, J.: Normalized loss functions for deep learning with noisy labels. In: ICML, pp. 6543–6553 (2020)
20. Miyamoto, H.K., Meneghetti, F.C., Costa, S.I.: The Fisher-Rao loss for learning under label noise. Inf. Geometry 1–20 (2022)
21. Nagarajan, B., Marques, R., Mejia, M., Radeva, P.: Class-conditional importance weighting for deep learning with noisy labels. In: VISIGRAPP (5: VISAPP), pp. 679–686 (2022)
22. Nishi, K., Ding, Y., Rich, A., Hollerer, T.: Augmentation strategies for learning with noisy labels. In: CVPR, pp. 8022–8031 (2021)
23. Northcutt, C., Jiang, L., Chuang, I.: Confident learning: estimating uncertainty in dataset labels. J. Artif. Intell. Res. **70**, 1373–1411 (2021)
24. Oyen, D., Kucer, M., Hengartner, N., Singh, H.S.: Robustness to label noise depends on the shape of the noise distribution in feature space. arXiv preprint arXiv:2206.01106 (2022)
25. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: a loss correction approach. In: CVPR, pp. 1944–1952 (2017)
26. Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G.: Learning from noisy labels with deep neural networks: a survey. IEEE Tran. NNLS (2022)
27. Sun, Z., et al.: PNP: robust learning from noisy labels by probabilistic noise prediction. In: CVPR, pp. 5311–5320 (2022)
28. Valle-Pérez, G., Camargo, C.Q., Louis, A.A.: Deep learning generalizes because the parameter-function map is biased towards simple functions. arXiv e-prints arXiv:1805.08522 (2018)
29. Wang, H., Xiao, R., Dong, Y., Feng, L., Zhao, J.: ProMix: combating label noise via maximizing clean sample utility. arXiv preprint arXiv:2207.10276 (2022)
30. Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G., Liu, Y.: Learning with noisy labels revisited: a study using real-world human annotations. arXiv preprint arXiv:2110.12088 (2021)
31. Wu, P., Zheng, S., Goswami, M., Metaxas, D., Chen, C.: A topological filter for learning with label noise. In: NIPS, vol. 33, pp. 21382–21393 (2020)
32. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. ACM **64**(3), 107–115 (2021)
33. Zhang, Y., Niu, G., Sugiyama, M.: Learning noise transition matrix from only noisy labels via total variation regularization. In: ICML, pp. 12501–12512 (2021)
34. Zheltonozhskii, E., Baskin, C., Mendelson, A., Bronstein, A.M., Litany, O.: Contrast to divide: self-supervised pre-training for learning with noisy labels. In: WACV, pp. 1657–1667 (2022)
35. Zhou, X., Liu, X., Zhai, D., Jiang, J., Ji, X.: Asymmetric loss functions for noise-tolerant learning: theory and applications. IEEE Trans. PAMI (2023)